上海工程技术大学

（勤奋、求是、创新、奉献）

2022～ 2023 学年第一学期考试试卷

主考教师：熊玉洁

学院 _____ 班级 _____ 姓名 _____学号 _____

《大数据基础平台》课程试卷 A

（本卷考试时间 120 分钟）

题号	一	二	三	四	五	六	七	八	九	十	总得分
题分	10	20	20	25	25						100
得分											

一、 判断题（本题共 10 小题，每小题 1 分，共 10 分）

题号	1	2	3	4	5	6	7	8	9	10	得分
√/×											

1. 大数据平台 Hadoop 的最重要的两个核心组件是 HDFS 和 Hbase。（ ）
2. 大数据场景下，数据的价值密度远远高于传统型数据的数据价值密度。（ ）
3. 大数据计算模式包括批处理计算、流计算、图计算以及查询分析计算。（ ）
4. Hadoop 基于 C++开发，具有较好的计算效率和性能。（ ）
5. HDFS 的局限包括不支持低延迟访问、多用户写入和任意文件修改。（ ）
6. HDFS 名称节点中 EditLog 记录了针对文件的所有更新操作。（ ）
7. HBase 为每个 Region 服务器维护一个 HLog 文件。（ ）
8. Hbase 中客户端每次获取 Region 位置，必须都经过三级寻址。（ ）
9. NoSQL 需要严格遵循 ACID 约束。（ ）
10. Hadoop2.0 设计了 HDFS HA，可以实现名称节点的冷热备份。（ ）

二、 填空题（本题共 10 个小题，每题 2 分，共 20 分）

1. 第三次信息化的浪潮主要发生在 2010 年前后，其标志是物联网、_____和大数据，可用于解决信息爆炸的问题。
2. 1ZB 等于_____GB（计算出具体数值）。

3. 物联网可以依照功能分为：感知层、网络层、处理层和_____。
4. HDFS 采用了块的概念，它带来的好处包括：_____、简化系统设计和适用数据备份。
5. HBase 通过_____可以避免单点失效的问题。
6. YARN 是一个纯粹的_____框架，可以在上面运行不同的计算模型。
7. 云计算类型主要包括_____、Platform as a Service 和 Software as a Service（单词全拼）。
8. 关系型数据无法满足 Web2.0 的需求，主要体现在_____、无法满足高并发的需求和无法满足高可扩展性和可用性的需求。
9. MapReduce 的核心环节的核心思想可以用_____来描述。
10. HDFS 联邦中的命名节点提供了_____和块管理功能。

三、 单项选择题（本题共 10 个小题，每题 2 分，共 20 分）

1. 大数据的计算模式包括（　　）。

   A) 批处理计算　　B) 流计算　　C) 查询分析计算　　D)以上皆是

2. 科学研究经历的四个范式，其中第三范式为（　　）。

   A) 实验科学　　B) 理论科学　　C) 计算科学　　D) 数据科学

3. 以下（　　）不是 HDFS 要实现的目标。

   A) 兼容廉价硬件设备　　B)简单文件模型

   C) 低延时数据访问　　D)大数据集

4. Hbase 的最核心的模块是（　　）。

   A) Region 服务器　　B)HDFS

   C) Zookeeper　　D)Master 服务器

5. 以下 NoSQL 数据类别和典型代表对应错误的是（　　）。

   A) 键值数据库----Redis　　B) 键值数据库----BigTable

   C) 文档数据库----MongoDB　　D) 图形数据库---GraphX

6. 以下关于 MapReduce 的描述错误的是（　　）。

   A) 用户不可以显式地从一台机器向另一台机器发送信息

   B) Map 任务之间不可以通信，但不同的 Reduce 任务之间可以进行信息交换

   C) 执行过程中，Map 任务的输入和 Reduce 任务的输出结果都保存在分布式文件系统里，而 Map 任务得到中中间结果保存在本地储存

   D) 以上皆是

7. Shuffle 是 MapReduce 的核心环节，以下描述正确的是 (　　)。

A) Shuffle 过程包括在 Map 端的 Shuffle 过程和在 Reduce 端的 Shuffle 过程

B) Map 端的 Shuffle 过程中，Map 任务的结果会直接写入磁盘

C) Reduce 端的 Shuffle 过程相对 Map 端而言，要更加复杂

D) 以上皆错

8. 以下 Spark 支持部署模式包括(　　)。

A) Standalone 模式　　B) Spark on Mesos 模式　　C) Spark on YARN 模式　　D)　以上皆是

9. YARN 相对 MapReduce1.0 的优势在于：(　　)。

A) 集中管理，增加了 ResourceManager 中心服务的功能和资源

B) 规定了标准化的 ApplicationMaster，要求不同计算框架使用相同的 ApplicationMaster

C) 使用容器为单位进行资源调度和分配，避免资源闲置浪费

D) 以上皆是

10. 以下属于监督学习方法的有（　　）。

A) K-means 聚类

B) 线性判别分析

C) 主成分分析

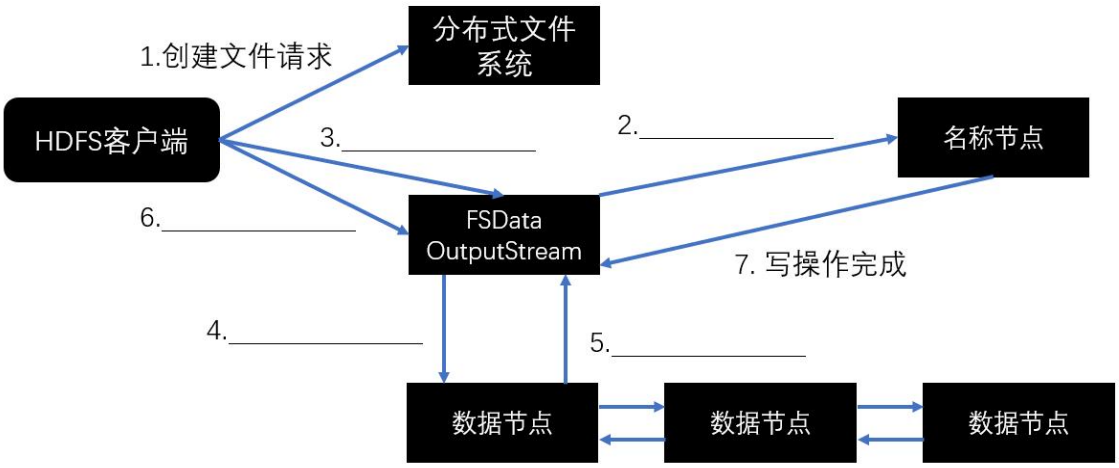
D) 关联分析

四、 简答题（本题共 5 个小题，每题 5 分，共 25 分）

1. 写出信息科技为大数据时代提供三个核心技术支撑（3 分），并进行简述它们三者的联系（2 分）。

2. 简述大数据技术在五个不同行业的应用示例（5 分）。

3. 补充下图中客户端向 HDFS 写数据的工作过程（5 分）。



4. 简述 Hbase 系统架构中的重要组成部分（4 分），并介绍 HBase 和 HDFS 的关系（1 分）。

5. 简述协同过滤算法的典型代表（2 分），并选择其中一种介绍其算法思想（3 分）。

五、 综合题（本题共 3 个小题，共 25 分）

1. 根据要求，给出合适的命令，实现 HDFS 的常用操作（每条命令 2 分，共 4 分）。

A．这里将/opt/data/SEEE 本地文件夹中的所有文件复制到 HDFS 的新建的 SUES 文件夹中。
命令 1：_____
命令 2：_____

2. 根据要求，给出合适的命令，实现 YARN 常用操作（每条命令 2 分，共 6 分）。

A．提交应用：使用 jar 命令提交一个 MapReduce 程序，其中程序代码位于/usr/SUES/hadoop-mapreduce-examples.jar，要求使用 Map 任务数目为 20，Map 任务的样本总数为 1000。
命令：_____

B. 使用 pyspark 命令启动一个 Spark 应用。并在此基础上使用 yarn application -list 命令，结合选项 appStates 列出所有状态为已提交的应用。
命令 1：_____
命令 2：_____

3. 结合个人理解，介绍当前大数据仍未完全得到应用的两个领域（2 分），说明可能的原因（3 分）。

4. 介绍 RDD 的概念（3 分），对应的依赖关系类型及表现形式（2 分）；写出 Spark 在划分依照 RDD 依赖关系进行阶段划分的原则（2 分），并依照此原则划分下图的不同阶段（3 分）。